David Parsons

dparsons@syr.edu

Final Project

Soccer Transfer Market Analysis

ARM, Clustering, Naïve Bayes, and SVM

9/20/2022

|  |  |
| --- | --- |
| Soccer Transfer Market Analysis | |
| **Introduction**10/10 | Currently the world soccer market is worth over 15b annually. There are leagues all around the world competing for various trophies, titles, and promotions. With so much on the line every club is looking to improve their club as much as possible, and quickly.  There are two main ways that these clubs can get better players and improve. A club can sign a player when they are a kid and couch them up through multiple youth clubs and leagues. This can take years to return results as well as has the risk of the prospect not developing as planned. Then there are transfers.  Transfers are a much more common and faster way to get a club to improve. A transfer is the way players move from club to club. It is when one club pays a fee to be able to sign a player that is under contract with a club already. A club from one league can do this with any league around the world, not just their own.  Soccer transfer updates: summary 8 July 2021 - AS USA  With so much money available in competitive market it is important to understand it and find out the best ways to approach it. There are new owners of these clubs all they time and each one needs to figure out this transfer market as soon as possible.  One of those new owners has reached out to a consulting firm to help give them a breakdown on the transfer market. This analysis will be used by the new club owner and front office to help them make transfers during the upcoming season. |
| **Analysis**40/40 | data preparation and cleaning The data set was obtained from Kaggle.com with two original files. The files were identical except for an extra column in the second file that had a Boolean value of is the move was a loan or not. For extra analysis the second file with this value was used in all analysis.  The overall data set had 2000 instances of transfers from leagues from all around the world. Each instance had data going over the player that was moved, their original club info, their new club info, and market values of the transaction.   data exploration The initial exploration showed that the majority of the data set was nominal variables with two numeric values.  Text, letter  Description automatically generated  For the two numeric values the summary showed that the market\_value of players was higher than the fees that the players went for.  Text  Description automatically generated with low confidenceText  Description automatically generated  Most of the analysis needed factorized data so the was converted. This helped show how there was moves from 82 different league, 556 clubs, and 60 countries.    The market\_value and fee columns were discretized into buckets to help with the models and analysis that needed factor data types as well as help with the majority of the fees being on the lower end of the range. models and methods The analysis was broken into three main sections: ARM, Naïve Bayes and SVM predictive models, and clustering using Kmeans.  With the data set being a list of transactions ARMs will provide an idea of if there are any similarities in these transactions.  Both naïve bayes and SVM models will be used to help predict the country\_to and fee of the transfers. For naïve bayes the naïve\_bayes function from the naivebayes library will be used.  Kmeans clustering will be used to analyze how closely related certain transfers are. This allow for a good visualization of how similar or dissimilar the transfers are. analyses GOALS and Parameters For the ARM analysis the goal is to find the top rules for the overall market, any interesting league specific rules, and rules for star players that are moving clubs. For the main rules there will be thresholds of supp = 0.005, conf = 0.90, minlen = 3, maxlen = 10 to get a decent size. For all other rule analysis, the support was lowered to .003 while the other criteria remained the same.  The naïve bayes and SVM models were used to predict the transfer fees of transactions as well as predict the league a player will go to. All four models used an 80/20 split of training and testing data with the same variables in both. The SVM used a polynomial kernel with a cost of 3.  Kmeans clustering was used to group together similar transfers with a focus on leagues and money. To do this the country and club data was removed. To get a cluster with both the numeric and nominal data the gower.dist function was used from the StatMatch library. This returned the distance (dissimilarity) of the nominal values. Both Euclidean and Manhattan methods were used for Kmeans. |
| **results**40/40 | technical results ARM:  For the top rules for the market the observation was transfers within leagues/countries were the most common. 13 of the 17 most common rules had some variation on this. The other top rules showed that players worth between 10m-50m were going to the Premier League and Italy had a number of transfers of players in the 1m-2.5m and 2.5-5m buckets.  Graphical user interface, table  Description automatically generated  When removing some country specific data to get more rules of the market there were a few noticeable ones. The club Nottingham Forrest had the most moves of players worth 10m-50m. It also showed that the most highly valued players were aged 26 or played left winger.  Graphical user interface, table  Description automatically generated  Looking at the Premier League specific rules showed that they had multiple clubs that were getting a noticeable number of players for 10m-50m. The second most common rule showed that the league that sold players to the Premier League the most was the Championship (the league beneath the Premier League).  Graphical user interface, application  Description automatically generated  The Italian league, Serie A, had a larger number of rules than any other league. It showed that Serie A clubs mainly bought players from with Serie A. Most of the moves in Serie A were under 5m except for two clubs; Juventus and Napoli. Two more interesting rule appeared regarding two clubs buying players in bulk. US Salernitana bought six centre-forwards and Udinese Calcio bought six left backs. This is twice the amount that a club normally has in on a roster.  The star player analysis showed that there were only two clubs that had a noticeable number of players brought in, West Ham and Nottingham Forrest. It also showed that the Premier League brought in seven star left wingers.    Naïve Bayes and SVM:  For the transfer fee prediction using the naïve bayes and SVM models there were vastly different results. The first picture shows the results of the naïve bayes model and second shows the SVM results. The SVM had a 99% accuracy while the naïve bayes only reached 43%.  Table  Description automatically generated Table  Description automatically generated  The prediction models for the league transferred to improved for both naïve bayes and SVM. For the naïve bayes model the accuracy improved to 99% and the SVM was up to 100% accurate.Table  Description automatically generatedTable  Description automatically generated  Kmeans Clustering:  For the star player (players with fees over 20m) clustering there were 4 distinct clusters. The green in the top left showed the most expensive players that transferred to the Premier League. The bottom left teal had the less expensive players going to the Premier League. The top right purple had the higher priced players that went to clubs outside the Premier League. Lastly the bottom right red cluster was made up of the lower priced transfers outside the Premier League.  A picture containing graphical user interface  Description automatically generated  Since the analysis was heavily skewed to either league\_to = Premier League or not a deeper look into the Premier League transfers was done.  Similarly, to the star transfers price and country were the main differences in the clusters. The top half had three clusters, red, blue, and yellow. These were all players that came to the Premier League from an outside league. The clusters decreased in transfer fee value going from left to right (red > blue > yellow). On the bottom were all the transfers of players changing clubs within the Premier League. The purple cluster on the left had the higher fee players when compared to the green cluster.  A picture containing timeline  Description automatically generated |
| **conclusions**10/10 | Conslusion The majority of clubs tend to make move within their league and country. There are benefits to looking at players from outside the league that a club is in. This could provide a larger pool of players that could help the team.  The Premier League has been the biggest league in the world for the past few decades and it shows in the transfer market. The transfer with those clubs tends to have a higher fee. This could price smaller clubs from other leagues out of the market.  While it is a large and market with many moving pieces there are ways to predict the transfer fees as well as the league a player will go to. This will help a club determine if they might be able to get a player to come to their league or if they can afford that player.  The transfers can be easily grouped together to help determine who is getting which players and for how much. This can help a club find out where they fall in the market and what tier of player might be able to come to their club.  With all this info it is a good start for a new club owner to get an understanding of the market. Transfers happen multiple times a year, so this is just the beginning of learning about the transfer market. As well as other data being available for a more detailed club and player level more analysis can be done in the future to better home in on transfer targets. All of this will hopefully get a club to a trophy winning level in no time.  Manchester City's first Premier League winning team - Where are they now? |